An Iterative Improved k-means Clustering

Madhuri A. Dalal[1] Nareshkumar D. Harale [2] Umesh L.Kulkarni [3]

[1]Student of Computer , M.G.M's college of Engineering, Kamothe, Navi Mumbai, India,madhuri_dalal25@yahoo.co.in
[2]Faculty of computer, M.G.M's college of Engineering, Kamothe, Navi Mumbai, India,nareshkumar.d.harale@gmail.com
[3]Faculty of Information, Konkan Gyanpeeth College of Engineering,Karjat,India,kumeshl@rediffmail.com

Abstract: **Clustering is a data mining (machine learning), unsupervised learning technique used to place data elements into related groups without advance knowledge of the group definitions. One of the most popular and widely studied clustering methods that minimize the clustering error for points in Euclidean space is called K-means clustering. However, the k-means method converges to one of many local minima, and it is known that the final results depend on the initial starting points (means). In this research paper, we have introduced and tested an improved algorithm to start the k-means with good starting points (means). The good initial starting points allow k-means to converge to a better local minimum; also the numbers of iteration over the full dataset are being decreased. Experimental results show that initial starting points lead to good solution reducing the number of iterations to form a cluster.**

Keywords: **data mining , clustering, k-means clustering, clustering algorithms.**

I. INTRODUCTION

Clustering is a data mining technique that separates your data into groups whose members belong together. This is similar to assigning animals and plants into families where the members are alike. Clustering does not require a prior knowledge of the groups that are formed and the members who must belong to it [17, 5]. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. From a practical perspective, clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others[1,17]. Data clustering is under vigorous development and is applied to many application areas including business, biology, medicine, chemistry, data mining and knowledge discovery [4], [7], data compression and vector quantization [5], pattern recognition and pattern classification [8], neural networks, artificial intelligence, and statistics.etc. Owing to the huge amounts of data collected in databases, cluster analysis has recently become a highly active topic in data mining research [2].The research has focused on finding efficient and effective cluster analysis in large databases. Classifying objects according to similarities is the base for much of science. Organizing objects into sensible grouping is one of the most fundamental modes of understanding and learning. Cluster analysis is the study of algorithms for grouping or classifying objects [8]. So a cluster is comprised of number of similar objects collected or grouped together[5][7].

There are two goals of clustering algorithms:

(1) Determining good clusters and

(2) Doing so efficiently.

Clustering is particularly applied when there is a need to partition the instances into natural groups, but predicting the class of objects is almost impossible. There are a large number of approaches to the clustering problem, including optimization based models that employ mathematical programming for developing efficient and meaningful clustering schemes. It has been widely emphasized that clustering and optimization may help each other in several aspects, leading to better methods and algorithms with increased accuracy and efficiency. Exact and heuristic mathematical programming based clustering algorithms have been proposed in recent years. However, most of these algorithms suffer from scalability as the size and the dimension of the data set increases.. Another important discussion in clustering is the definition of best partitioning of a data set, which is difficult to predict since it is a relative and subjective topic. Different models may result in different solutions subject to the selected clustering criteria and the developed clustering model [4][6][13]. For cluster analysis to work efficiently and effectively, as

many literatures have presented, there are following typical requirements of clustering in data mining:

1. Scalability:

That is to say an efficient and effective clustering method should not only be able to work well on small data sets, but also a large database containing about millions of objects.

2. Ability to deal with different types of attributes:

An efficient and effective clustering method is required to cluster various types of data, not only numerical, but also binary, categorical, and ordinal data, or mixtures of these data types.

3. Discovery of clusters with arbitrary shape**:**

A cluster could be of any shape. It is important to develop algorithms that can detect clusters of arbitrary shape.

4. Minimal requirements for domain knowledge to determine input parameters**:**

Some algorithms require users to input certain parameters in cluster analysis. But the parameters are hard to determine.

5· Ability to deal with noisy data:

A good clustering algorithm is required to be independent of the influence of noise.

6· Insensitivity to the order of input records:

DOI: 01.IJNS.02.03.183

It is important to develop algorithms that are insensitive to the order of input.

7· High dimensionality:

The ability to handle high-dimensional data is important for a good algorithm.

Several clustering algorithms have been proposed. These algorithms can be broadly classified into hierarchical and partitioning clustering algorithms [8]. Hierarchical algorithms decompose a database *D* of *n* objects into several levels of nested partitioning (clustering), represented by a dendrogram (tree). There are two types of hierarchical algorithms; an agglomerative that builds the tree from the leaf nodes up, whereas a divisive builds the tree from the top down. Partitioning algorithms construct a single partition of a database *D* of *n* objects into a set of *k* clusters, such that the objects in a cluster are more similar to each other than to objects in different clusters. The k-means clustering algorithm is the most commonly used partitioned algorithm [8][12] because it can be easily implemented, speed convergence to local minimum. However this local minimum depends on the initial starting means. In this paper, we introduce an efficient improved method to obtain good initial starting means, so the final result will be better than that of randomly selected initial starting means. How to get good initial starting means becomes an important operational objective [1][14]. This paper is organized as follows: Section 2 introduces K-means clustering and our proposed improved iterative k-means clustering method. Section 3 presents experimental results comparing both algorithms. Section 4 concludes the paper.

II . K-MEANS CLUSTERING

A. *k-means Clustering:*

There are many algorithms for clustering datasets. The k-means clustering is the most popular method used to divide *n* patterns {x1, …, xn} in *d* dimensional space into *k* clusters [8]. The result is a set of *k* centers, each of which is located at the centroid of the partitioned dataset. This algorithm can be summarized in the following steps:

1. Initialization: Select a set of *k* starting points {*mj*}, *j* = 1, 2… *k*. The selection may be done in random manner or according to some heuristic.

2. Distance calculation: For each pattern xi, 1<=i<=n, compute its Euclidean distance to each cluster centroid mj, 1<=j<=k , and then find the closest cluster centroid mj and assign the object *xi* to it.

3. Centroid recalculation: For each cluster j, 1<=j<=k, recomputed cluster centroid mj as the average of the data points assigned to it.

4. Convergence condition: Repeat steps 2 and 3 until convergence.

To choose a proper number of clusters *k* is a domain dependent problem. To resolve this, some researchers have proposed methods to perform k-clustering for various numbers of clusters and employ certain criteria for selecting the most suitable value of *k* [15] and [16]. Several variants of the k-means algorithm have been proposed. Their purpose is to improve efficiency or find better clusters. Improved efficiency is usually accomplished by either reducing the number of iterations to reach final convergence or reducing the total number of distance calculations. The k-means algorithm randomly selects *k* initial cluster centers from the original dataset. Then, the algorithm will converge to the actual cluster centers after several iterations. Therefore, choosing a good set of initial cluster centers is very important for the algorithm. However, it is difficult to select a good set of initial cluster centers randomly.

B. *Iterative Improved k-means Clustering:*

In this section we describe our algorithm that produces good starting points for the k-means algorithm instead of selecting them randomly. And this will leads to better clusters at the final result than that of the original k-means. In the model in this study, it is assumed that the number of desired clusters k is known a priori since the determination of the number of clusters constitutes another subject of research in the clustering literature. However, typically k will be small. The goal of the model is to find the optimal partitioning of the data set into *K* exclusive clusters given a data set of *n* data items in m dimensions, i.e. a set of n points in Rm. The parameter d_{ij} denotes the distance between two data points i and j in R_m and can be calculated by any desired norm on R_m such as the Euclidean distance which is is the straight-line distance with two points. The Euclidean distance between point's p and q is the length of the line segment. In Cartesian coordinates, if p = (p_1, p_2,..., p_n) and q = (q_1, q_2... q_n) are two points in Euclidean *n*-space, then the distance from p to **q** is given by:

$$\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_n-q_n)^2}=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}.$$

Consider a data set D = {d(j) = ($d_1^{(j)}$, $d_2^{(j)}$,…….. $dm^{(j)}$ | j = 1…….n } in R_m and K be predefined number of clusters. Bellow is the outline of a precise cluster centers initialization method. K

Step1 : Dividing D into K parts as D =U $S_{k,}$ S_{k1}) ∩ S_{k2} = $= \emptyset, k1 \neq k2$ k=1 according to data patterns.

Step2 : Calculate new c_k centers as the optimal solution of

$$\min_z = \sum_{d(j)\,\varepsilon\,S_k} x - d^{(j)} \| \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

x = (x_1. . . x_m) ε Rm where || • || denotes the 2-norm.

Step 3 : Decide membership of the patterns in each one of the K-clusters according to the minimum distance from cluster center criteria.

Step 4 **:** Repeat steps 2 and 3 till there is no change in cluster centers.

The step 1 is rather flexible. We can accomplish it according to the visual inspection or any other methods. Normally this heuristic partition is better than random sampling. Also we can use sub-algorithm below to accomplish step 1 for any data patterns.

C. *Sub-algorithm:*

1) Compute the d_{min} and d_{max} by

$$d_{min} = \min_{1 \le j \le n} \| d(j) \|$$

DOI: 01.IJNS.02.03.183

$$d_{max} = \max_{1 \le j \le n} \| d(j) \|$$

1) For k = 0, 1, . . . , *K* "1, calculate new c_k centers as:

$c_k = d_{min} + ((k/K) + 1/(2* K))(d_{max} - d_{min})$ …..(2)

Step 2 in sub algorithm has been modified for one-dimensional dataset as follows :

$$c_k = d_{min} + ((k/2*k + (5.5/(4*k))(d_{max} - d_{min}) \text{…….(3)}$$

Theoretically, it provides us an ideal cluster center. Nevertheless, the process of finding the optimal solution to problem (3) is too expensive if the problem is computed by normal optimization algorithm .Fortunately we can establish a simple recursive method to handle problem (1). We propose a novel iterative method which is proved to be with high efficiency by our simulation results.

We obtain the iterative formula

$$x_i^{(k+1)} = \left(\sum_{d^{(j)} \in S_k} d_i^{(j)} / q_i^k\right) / \left(\sum_{i=1}^{n} 1/q_i^k\right) \quad \text{..... (4)}$$

Where $q_j^k = \| x^{(k)} - d^{(j)} \|$

The above iteration process requires an initial point as its input. We can simply use the average of the coordinates of the data points.

$$x^{(k)}_{(0)} = \sum_{d(j) \in S_k} d(j) / |S_k| \quad \text{.....(5)}$$

k= 1,2 ,3 ……..K

D. Iterative improved k-means clustering Algorithm:

1. Dividing D into K parts as $D = \cup S_k$, $Sk1) \cap Sk2 = \emptyset, k1 \neq k2$. k=1 according to data patterns.(call sub algorithm)
2. Decide membership of the patterns in each one of the K-clusters according to the minimum distance from cluster center criteria
3. Calculate new centers by iterative formula (5).
4. Repeat steps 3 and 4 till there is no change in cluster centers.

III. EXPERIMENTAL RESULTS

We have evaluated our proposed algorithm on Fisher's iris datasets, Pima Indian medical diabetes dataset, and soya bean plant dataset considering one-dimensional data as well as multi-dimensional data. We compared our results with that of the original k-means algorithm in terms of the number of iterations for both algorithms. We give a brief description of the datasets used in our algorithm evaluation. Table 1 shows some characteristics of these datasets [9][10][11].

TABLE 1: CHARACTERISTICS OF DATASETS

Dataset	No of records (N)	No of attributes
Fisher's Iris dataset	150	4
Pima Indian Medical diabetes	768	8
Soya bean Plant dataset	47	35

A. One dimensional Dataset:

To gain some idea of the numerical behavior of the improved k-means algorithm and to compare it with the original K-means algorithm of randomly choosing initial starting points, we first solve a problem in detail by original and improved k-means algorithm, with the same dataset separately. The cardinality of data set is given by column labeled N in table 1. Total number of iterations, required for entire solution of each dataset is displayed in two columns labeled k-means and improved k-means under iterations. The following table shows the number of iterations taken by k-means and improved k-means for k= 3 considering 1-D dataset.

TABLE II: COMPARISON OF K-MEANS AND IMPROVED K-MEANS FOR

Dataset	Iterations	
	K-means	*Improved k-means*
Fisher's Iris Dataset	8	2
Medical Diabetes Dataset	5	3
Soya bean Plant Dataset	3	2

Results for k-means and improved k-means can be plotted graphically for all datasets as shown below.

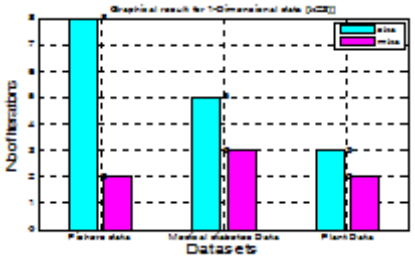

Fig. 1: Iteration comparison between k- means and improved k-means for k=3

B. Multi dimensional Dataset:

Table III: Comparison of k-means and improved k-means for k=3

Dataset	Iterations	
	K-means	*Improved k-means*
Fisher's Iris Dataset	12	10
Medical Diabetes Dataset	25	15
Soya bean Plant Dataset	4	3

Results for k-means and improved k-means can be plotted graphically for all datasets as shown below.

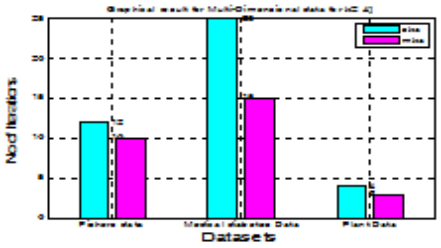

Fig 2: Iteration comparison between k- means and improved k-means for k= 3

TABLE IV: COMPARISON OF K-MEANS AND IMPROVED K-MEANS FOR K=4

Dataset	Iterations	
	K-means	*Improved k-means*
Fisher's Iris Dataset	13	7
Medical Diabetes Dataset	38	30
Soya bean Plant Dataset	4	3

DOI: 01.IJNS.02.03.183

Results for k-means and improved k-means can be plotted graphically for all datasets as shown below.

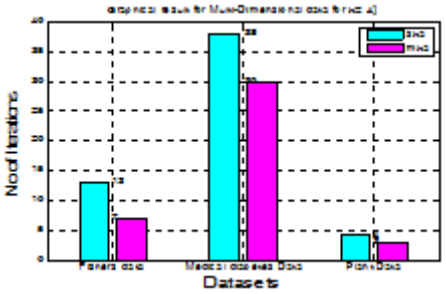

Fig. 3: Iteration comparison between k- means and improved k-means for k= 4

It is seen that our experimental results shows that number of iterations required by improved k-means are less as compared to that of original k-means algorithm and the margin between the total number of iterations required by k-means and improved k-means is much larger.

IV. CONCLUSION

This paper presented iterative improved k-means clustering algorithm that makes the k-means more efficient and produce good quality clusters. We analyze the solutions of two algorithms namely original k-means and our proposed method, iterative improved k-means clustering algorithm .Our idea depends on the good selection of the starting points for the k-means. This is based on the optimization formulation and a novel iterative method. It can be applied to many different kinds of clustering problems or combined with some other data mining techniques for getting more promising results. The experimental results using the proposed algorithm with different datasets are very promising. It is seen that iterations required by iterative improved k-means algorithm are fewer than those original k-means algorithm. Our experimental results demonstrated that the proposed algorithm produces better results than that of the k-means algorithm

ACKNOWLEDGMENTS

I thank Prof. Umesh L. Kulkarni and Prof. Naresh. D. Harale for their assistance with the implementation of the algorithms and for many useful and interesting discussions. I also thank UCI machine learning repository for providing the original dataset.

REFERENCES

[1] Bradley P.S., Fayyad U.M., "Refining Initial Points for K-Means Clustering", Proc. of the 15th International Conference on Machine Learning (ICML98), J. Shavlik (ed.), Morgan Kaufmann, San Francisco, 1998, pp. 91-99.

[2] Zhijie Xu,Laisheng Wang , Jiancheng Luo ,Jianqin Zhang "Modified clustering algorithm for data mining " 2005, 741-744

[3] Sugar S, James G, " Finding the number of clusters in a dataset : an information theoretic approach. " Stat.Assoc. 98(2003) 750-763

[4] Su,Chou C, "A modified version of the K-means algorithm with a distance based on cluster symmetry." IEEE Transaction Pattern Analysis Machine Intel , 23(6) (2001) 674-680

[5] M.N. Mutty, A.K. Jain , P.J.Flynn , "Data clustering : a review", ACM computing serveys 31(3) (1999)264-322

[6] Shehroz S. Khan, Amir Ahmad, "Cluster center initialization algorithm for K-means clustering." , Pattern Recognition Letters 25 (2004) 1293-1302 Operational Research, 174(2)(2006) 930-944

[7] A.K. Jain, R.C. Dubes, "Algorithms for Clustering Data" , Prentice Hall, Englewood Cliffs, NJ(1988)

[8] Wei Li ,"Modified k-means Clustering algorithm" ,congress on image and signal processing, 2008,618-621

[9] Fishers Iris Dataset
http://www.math.uah.edu/stat/data/Fisher.txt

[10] Pima-indians Medical Diabetes dataset
http://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.data

[11] Soyabean plant small dataset
http://archive.ics.uci.edu/ml/machine-learning-databases/soybean/

[12] Fahim A. M., Salem A. M., Torkey F. A. and Ramadan M. "An efficient enhanced k-means clustering algorithm". Journal of Zhejiang University Science A, 2006,vol7(10),pp1626-1633.

[13] Joaquín Pérez Ortega1 , Ma. Del Rocío Boone Rojas,1,2, María J. Somodevilla García2 ,Research issues on K-means Algorithm: An Experimental Trial Using Matlab.

[14] Meila, M., Heckerman, D., An experimental comparison of several clustering methods, Microsoft Research Report MSR-TR-98-06, Redmond, WA.(1998)

[15] Pham D. T., Dimov S. S., and Nguyen C. D., "Selection of *k* in K- means clustering". Mechanical Engineering Science, 2004, vol. 219. pp.103-119.

[16] Ray S. and Turi R. H., "Determination of number of clusters in k- means clustering and application in colour image segmentation.", in Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques, 1999, pp. 137-143,

[17] J M Pena, J A Lozano, P Larranaga, "An empirical comparison of four initialia-zation methods for the K-means algorithm." Pattern Recognition Lett. 20(1999) 1027-1040

DOI: 01.IJNS.02.03.183